KANT: KNOWLEDGE-BASED, ACCURATE NATURAL LANGUAGE TRANSLATION

Teruko Mitamura, Eric Nyberg, Jaime Carbonell
Center for Machine Translation
Carnegie Mellon University

KANT is an interlingual MT system for multi-lingual translation of technical documents, written using a controlled vocabulary and grammar.

KANT is comprised of a set of software modules (parser, interpreter, mapper, generator) which work together to produce target language translations from controlled source text. These modules are the result of long-term research and development in practical machine translation at the Center for Machine Translation (CMT) at Carnegie Mellon University, located in Pittsburgh, PA. The KANT software grew out of extensions and refinements to earlier systems developed at the CMT, which include the CMT-SEMSYN system, a collaborative effort with the University of Stuttgart in the domain of doctor patient communications (Japanese and English source languages to Japanese, English and German target languages), and the KBMT-89 system, a funded project with IBM's Tokyo Research Laboratory in the domain of PC installation manuals (Japanese and English to Japanese and English; cf. (Goodman and Nirenburg, 1991)).

1 The KANT Approach

KANT is a knowledge-based translation system. For each source language to be analyzed and each target language to be produced, the system makes use of specific lexicons, grammars, semantic rules, etc. to perform its task. The KANT software itself is language-independent — the same code modules are used regardless of the source and target languages. Extending the system to a new language involves writing a new lexicon, grammar, etc. for the language, but does not require any modification to KANT itself.

KANT is an interlingua-based translation system. For each sentence in the source language, KANT produces a semantic representation or "interlingua" expression. The interlingua is independent of the source and target languages, and is based on the set of concepts (objects, events, properties) relevant to the domain of translation. The generation phase in KANT produces a target language sentence for each interlingua expression. Because interlingua is used as an intermediate representation, the analysis of the source language and the generation of the target language are independent

of each other, which eliminates the need for bilingual dictionaries and transfer grammars for specific language pairs (see Figure 1).

Figure 1: Interlingua Translation

2 The KANT Domains

KANT is specifically designed for multi-lingual document production from a single, controlled source language. KANT is intended for use in domains where technical information (such as documentation for machinery, electronics, software/hardware, etc.) is to be authored preferably at a central location and translated for dissemination world-wide. KANT achieves high accuracy (with no need for post-editing) precisely because it takes advantage of a controlled source language and a well-defined application domain.

The source language is not limited by the size of its vocabulary; rather, KANT places certain restrictions on the types of sentences and phrases that can be used, to eliminate unnecessary vagueness and ambiguity in the source text. This not only improves translatability, but also encourages uniform, concise text in the source language as well. Although many constructions are eliminated from the source language, the set of constructions supported by KANT is more than expressive enough for the authoring of technical information in a particular domain.

3 Ongoing Applications

The first KANT application, ESTRATO, translated English/Spanish texts in the domain of electric power utility management (Union Electrica Fenosa). The most prominent ongoing KANT application is for 11 languages in the heavy equipment domain, produced by the CMT and Carnegie Group for Caterpillar, Inc. Currently, the Controlled English, Language Environment (CGI), and French (CMT) modules have been delivered to Caterpillar.